Synch-Graph: Multisensory Emotion Recognition Through Neural Synchrony via Graph Convolutional Networks

Esma Mansouri-Benssassi and Juan Ye
University of St Andrews
School of Computer Science
St Andrews, United Kingdom
{emb24, juan.ye}@st-andrews.ac.uk

Abstract

Human emotions are essentially multisensory, where emotional states are conveyed through multiple modalities such as facial expression, body language, and non-verbal and verbal signals. Therefore having multimodal or multisensory learning is crucial for recognising emotions and interpreting social signals. Existing multisensory emotion recognition approaches focus on extracting features on each modality, while ignoring the importance of constant interaction and co-learning between modalities. In this paper, we present a novel bio-inspired approach based on neural synchrony in audio-visual multisensory integration in the brain, named *Synch-Graph*. We model multisensory interaction using spiking neural networks (SNN) and explore the use of Graph Convolutional Networks (GCN) to represent and learn neural synchrony patterns. We hypothesise that modelling interactions between modalities will improve the accuracy of emotion recognition. We have evaluated Synch-Graph on two state-of-the-art datasets and achieved an overall accuracy of 98.3% and 96.82%, which are significantly higher than the existing techniques.

Introduction

Human perceives emotions in a multisensory manner, where information from different sensory modalities such as facial expression, verbal and non-verbal signals, and body languages expresses our emotional states. The multisensory emotional precept is conveyed through a constant cross-talk between various sensory modalities. Understanding emotions from various modalities is crucial for human computer interaction in a multitude of applications such as gaming, mental health or car driving. Therefore it is important to translate the multisensory relationship between different modalities in order to get a better meaning and more accurate interpretation of emotions.

Research in multimodal emotion recognition mainly focuses on feature extraction in individual modalities and integration by applying state-of-the-art data fusion techniques previously derived from engineering (Baltrušaitis, Ahuja, and Morency 2018). Early feature fusion often involves extracting features from visual and auditory modalities; *e.g.*, using LSTM for temporal feature extraction from video, and concatenate features that are fed to a SVM for emotion recognition (Chao et al. 2016). Decision fusion techniques introduce another layer on top of inferences from each modality; *e.g.*, applying Dynamic Bayesian Network on decisions from parallel models on audio and visual data (Felipe, Luis J, and Pedro 2015).

Recently deep learning techniques have also been applied to fusion tasks, not only in feature extraction. For example, Zhang et al. have employed CNN and 3D-CNN to extract spatial and temporal features on audio and visual segments. The features are then fused into a deep belief network (DBN) model to learn discriminating global feature representations (Zhang et al. 2017).

Although these techniques produce promising results, most of them have not taken into account the constant cross-talk and temporal relationship between different modalities in multisensory emotion recognition tasks (Wagner and André 2018). The recent study in cognitive neuroscience on cross-modal modulation in emotion processing (Garrido-Vásquez et al. 2018) has shown that cross-modal interaction is particularly important in emotion recognition, where signals from different modalities can complement each other in learning and thus signals in one modality can used to predict the other. For example, dynamic facial expressions can influence vocal emotion processing.

Driven by this research problem, we propose a novel multisensory emotion recognition approach based on temporal neural synchrony and phase-coupling in the brain (Symons and et al 2016; Keil and Senkowski 2018).

We hypothesise that modelling neural synchrony and constant cross-talk between modalities will enhance the accuracy of emotion recognition. To the best of our knowledge, we are the first to investigate the use of graph neural networks to model bio-inspired neural synchrony; that is, learning synchronous patterns of neuron connections cross multiple modalities acquired from a bio-inspired architecture – spiking neural network (SNN).

SNN has demonstrated as a promising approach for capturing intrinsic features of visual and audio modalities for emotion recognition in a unsupervised learning manner (Mansouri-Benssassi and Ye 2019), where features are

represented in spiking patterns of neurons. Moving beyond, we employ SNN to capture constant connections between neurons in different modalities, which simulates how the brain perceives emotions from multiple sensory modalities. Learning patterns on these neuron connections is a challenging task due to the non-Euclidean nature of the data. Graph neural networks have been successfully applied to similar complex, unstructured data, such as chemical reactions and citation networks (Wu and et al 2019). Novelly we propose a new way to construct a graph network to model neuron interactions and employ graph convolutional network to learn connection patterns for multisensory emotion recognition.

The main novelty and contribution of our work is in three folds. Firstly, we explore the use of SNN in learning and representing interactions between signals in different modalities. Secondly we employ graph networks to model and learn interaction patterns for multisensory emotion recognition. Thirdly, our approach has been experimented and validated on two state-of-the-art datasets and has demonstrated consistently superior performance to the existing techniques.

Background on Graph Neural Network

Graph neural networks are gaining more and more attention in dealing with problems on unstructured data such as classification of social networks, representations of biological systems and chemical reactions. Following the success of Convolutional Neural Networks, Bruna et al. are one of the first who have applied convolutional layers to graph neural network (Bruna et al. 2013). They employ spectrum of the graph Laplacian that translates the convolutional properties into the Fourier domain. This results in a simpler representation of graph data.

Henaff et al. have applied graph convolutional network (GCN) and spectral learning to large classification problems such as ImageNet object recognition and bioinformatics (Henaff, Bruna, and LeCun 2015). They have designed unsupervised learning for graph estimation when the graph structure is unknown. To address the limitations of spectral methods for large graph, spatial convolutions are introduced, which allows learning functions by aggregating features between neighbouring nodes. They are particularly useful for node classification as they do not require to process the whole graph simultaneously as for spectral methods.

Kipf et al. have introduced a semi-supervised method using a localised first order approximation of spectral graph convolutions for node classification (Kipf and Welling 2016). This helps in alleviating the complexity challenge of spectral convolutions on processing whole graph. They experiment on citation networks and the results have shown that the model can effectively learn hidden layer representations encoding local graph structure and features of individual nodes.

Hamilton et al. overcome the challenge of large graphs by introducing inductive node embedding where node features are used to learn an embedding function generalising on unseen nodes. This is achieved by using the topological structure of local neighbours of each node. It trains on aggregator functions instead of feature vectors on each node. An unsupervised loss function is designed so as to enable training without using task-specific labels.

Gao et al. have used Learnable Graph Convolutional Layer (LGCL) to enable convolution operations on large graphs (Gao, Wang, and Ji 2018). This works by transforming the graphs into 1-D format grid to make the use of convolutions easier and more accurate.

They have developed subgraph training to reduce the computational complexity of the current training method that uses the whole adjacency matrix as an input.

Applications of GCN are starting to emerge in computer vision in general and affective computing recently. Nian et al. propose the use of GCN in facial features recognition (Nian et al. 2019). They have used GCN for defining facial attributes such as hair colour, eyes or brow shape. They first extract facial features using CNN, which are then transformed into the above attributes. These are used to construct a graph with nodes representing facial attributes and edges representing relations between them.

GCNs have been used for emotion recognition through EEG data (Song et al. 2018). Song et al. have proposed Dynamical Graph Convolutional Neural Network (DGCN) to model multi channel EEG features where each EEG channel represents a node in the graph. The adjacency matrix is learned in a dynamic way, where the matrix is updated at training time. This is the opposite of classical GCN where the adjacency matrix is often fixed at the beginning of the training.

Zhang et al. have used GCN to model context in emotion recognition (Zhang, Liang, and Ma 2019). They compute the relation between context information with a graph and one example of context is facial expression of the interlocutor. Then facial features are extracted with CNN and concatenated to context information.

GCNs have demonstrated promising results in various applications (Wu and et al 2019; Hamilton, Ying, and Leskovec 2017) and play an important role in the advancement of affective computing and emotion recognition. In this paper, we novelly apply GCN in modelling neural synchrony to learn complex interaction patterns between synchronised neuron activities captured in a spiking neural network. To the best of our knowledge, we are the first to apply GCN to bio-inspired multisensory emotion recognition.

Proposed Approach

This section describes a bio-inspired approach to model multisensory emotion recognition using neural synchrony. It consists of three main components: (1) simulating and modelling multisensory integration and interaction via spiking neural networks; (2) modelling neural synchrony through a graph network; and (3) applying graph convolution network to multimodal emotion recognition. In the following, we will describe each of these components in details.

Multisensory Integration and Interaction in Spiking Neural Network

Spiking neural networks simulate neuron activities in the brain. Information is transmitted between neurons using ac-

tion potentials via synapses in the brain. When a membrane potential reaches a certain threshold a spike is generated (Jose, Amudha, and Sanjay 2015). The computation of SNNs is based on timing of spikes rather than their shape. That is, spikes that fire together have a stronger connection. Neurons communicate through a series of spikes, which defines the unique patterns distinguishing different emotional states. To model the interaction between modalities, the SNN consists of three main layers:

1. An input layer receives unisensory signals in both visual and auditory modalities;
2. An excitatory layer comprising two excitatory neuron groups translates information from auditory and visual inputs into spike patterns;
3. An inhibitory layer with two neuron groups linked to the excitatory layer for each modality with a lateral inhibition; that is, a neuron in the inhibitory layer is connected to all neurons in the excitatory layer apart from the one it receives signal from.

In the following, we will describe the learning process in the SNN.

Neuron Activity Neurons in a SNN communicate through spikes, enabling them to learn specific features at the excitatory layer. Each neuron behaviour is modelled through Leaky-Integrate-and-Fire (LIF) (Diehl and Cook 2015), as defined in the following equation:

$$\tau \frac{dV}{dt} = (E_{rest} - V) + g_e(E_e - V) + g_i(E_i - V). \quad (1)$$

V is the membrane voltage and E_{rest} represents the membrane potential in the resting phase. E_i and E_e represent the equilibrium potential for both inhibitory and excitatory synapses. g_e and g_i is the conductance value of synapses at the excitatory and inhibitory layers.

Neurons fire when they reach a certain threshold. They then enter a resting phase E_{rest} for an interval of 5ms. At this moment neurons cannot spike as they are in a refractory phase. τ is a time constant representing the time a synapse reaches its potential. This is longer for excitatory neurons. This is set at 200ms and 100ms for excitatory and inhibitory neurons respectively. This delay is motivated by the learning process happening mainly at the excitatory layer. The choice of temporal parameters is not biologically realistic. This is justified by the number of input neurons being smaller than biological networks (Diehl and Cook 2015).

We have also applied *homeostasis* (Diehl and Cook 2015) through an adaptive membrane threshold V_{thresh} in order to have a more stable network and to refrain some neurons from spiking for all the inputs (Rathi and Roy 2018). At the inhibitory layer, all neurons are inhibited apart from the one they receive information, referred to as *lateral inhibition*. This is used to encourage competition between neurons.

Synapses conductance increases when pre-synaptic reaches the synapse before the post-synaptic otherwise they decrease exponentially. The dynamics is ruled by a time constant as defined in the following equation.

$$\tau_{g_e} \frac{dg_e}{dt} = -g_e \quad (2)$$

where τ_{g_e} is a time constant of post-synaptic potential. The time constant is set to 1ms for the inhibitory conductance and to 2ms for the excitatory conductance.

Unsupervised Learning Through STDP Learning in SNN is achieved in an unsupervised manner through Spike Timing Dependent Plasticity (STDP) (Diehl and Cook 2015). STDP has been successfully used in facial expression recognition (Benssassi et al. 2018) and speech emotion recognition tasks (Mansouri-Benssassi and Ye 2019). It is a form of Hebbian learning, where connections between neurons are created and strengthened when they fire at the same time. The main learning is influenced by the time of spiking of pre-synatptic and post-synaptic neurons. Weights are updated by the following equation:

$$\Delta w = \eta(x_{pre} - x_{tar})(w_{max} - w)^{\mu} \quad (3)$$

η is the learning rate. w_{max} is the maximum weight and x_{tar} is the target value of the pre-synaptic trace when the post-synaptic spike fires. This is used to enable the disconnection of neurons that seldom lead to firing, when the post-synaptic neuron is rarely active. μ is the dependence of updates on previous weight. x_{pre} is the pre-synaptic trace left every time pre-synaptic spike reaches a synapse. That is, weights are increased if pre-synaptic spikes fire prior to post-synaptic spikes. Otherwise, they decrease. The change of weights in STDP learning is computed by a function tracking differences in timing between pre-synaptic and post-synaptic spikes. STDP learning proves to be a simpler and advantageous method compared to classical supervised learning such as back-propagation (Hazan et al. 2018).

Modelling Neural Synchrony in Graph Network

To enable multisensory integration, we set recurrent connections at the excitatory layer between audio and visual neuron groups in order to allow cross-talk between modalities. This is achieved by connecting neurons that spike together between both modalities.

After training the SNN, we obtain information on neuron activities, including the location of a neuron at the excitatory layer, time of spiking in milliseconds, and the modality type of each neuron, which together defines patterns for each type of emotional states.

Neural synchrony represents neurons spiking within the same temporal window. This facilitates the integration of information from different sensory sources (Stein 2012); that is, learning and extracting relevant and crucial features from sensory inputs such as heterogeneous neuronal populations (Brette 2012). In this paper, we propose to model neural synchrony with a graph network. Neurons are modelled as nodes and their spiking synchrony as edges. In this way, we can learn complex patterns between visual and auditory neuron groups through graph neural network to enable multisensory emotion recognition.

We define a neural synchrony graph network as an undirected graph: $G = (V, E)$, where V is a set of nodes representing neurons and E defines edges of relations between nodes. The edges include two types of relations: temporal and stimuli based. Edges are added between nodes which spike within a temporal window of integration.

We also define node feature matrix $X_{N\times D}$, where N is the number of nodes and D is the dimension of input features on each node. Nodes represent neurons with features defining the type of neurons; that is, either audio or visual. Nodes are connected if they belong to the same video and spike within the temporal window of integration. We have set up the temporal window to 150ms to simulate a biologically realistic temporal window of integration in multisensory integration (Balconi and Carrera 2011).

We introduce an adjacency matrix A which describes the main structure of the graph network. The adjacency matrix is represented by a sparse matrix containing adjacency matrices for each subgraph that is constructed on a video input. The adjacency matrix is represented by two main aspects: *temporal coordination* between neuron spikes and *stimulus* based relations, where neurons belonging to the same subgraph and class type are linked together.

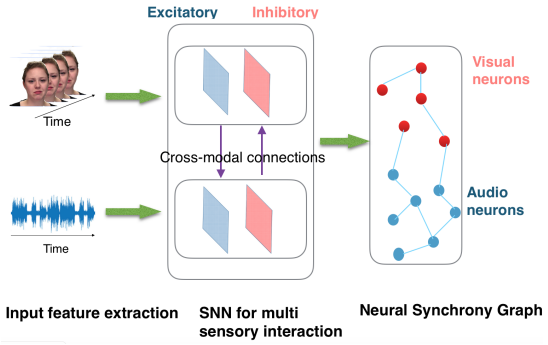

Figure 1: The workflow of our multisensory interaction graph modelling. First features are extracted from both visual and audio data, and then fed to a SNN where multisensory integration is simulated. After training, neuron activities are recorded, based on which a graph is constructed.

Multisensory Interaction Learning Workflow Figure 1 describes the process of multisensory integration and interaction in a SNN for graph construction. Given a video input, we segment it into visual frames and audio segments, from which we extract features. For visual features, we resize each frame and crop the face area, and then apply Laplacian of Gaussian (LoG) filters to extract contour and facial features. The LoG filter has been successfully applied for extracting high precision features (Mansouri-Benssassi and Ye 2018) and is represented in Equation 4.

$$\nabla^2 G_\sigma(x,y) = \frac{\partial G_\sigma(x,y)}{\partial x^2} + \frac{\partial G_\sigma(x,y)}{\partial y^2} \tag{4}$$

where ∇^2 is the Laplacian operator, σ is the smoothing value, and $G_\sigma(x,y)$ is the Gaussian filter applied to the image, given by:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

Gaussian filters are applied first to reduce noise, and then the Laplacian filter is applied to give a precise definition of contours, corners and facial features.

We extract MFCCs from each auditory segment. Both visual and auditory features are transformed into Poisson spike train as input to a SNN. Each modality will correspond to two different neuron groups connected at the excitatory layer. The inhibitory layer consists of two neurons groups each connected laterally to the excitatory neurons.

Given a video input, we obtain these two connected neuron groups, which will form into a subgraph. We compose each subgraph from videos in a complete graph, where neurons between subgraphs are connected if they share the same class label; *i.e.*, the same emotional state.

Multisensory Emotion Recognition via Graph Convolution Neural Network

We define emotion recognition as a subgraph classification problem; that is, assigning a class label to each subgraph. We build our architecture based on semi-supervised GCN model (Kipf and Welling 2016), which is applied to node classification in GCN. It employs layer wise propagation rule based on first-order approximation using spectral convolutions. Spectral convolutions represent filters as graph signal processing based on spectral theory. Introducing the first-order approximation (Kipf and Welling 2016) allows a simplification of the model and a faster training time. Their model is particularly useful for our neural synchrony multisensory emotion recognition model as it can better capture global complex patterns in graphs compared to spatial convolutions methods where they capture more local areas of nodes. Training the whole graph instead of node batches helps maintain the neural synchrony structure. The reason is that the classification of emotions is conveyed by the neural synchrony pattern instead of individual nodes.

We have adapted the model (Kipf and Welling 2016) for the subgraph classification by introducing an additional general pooling layer (Duvenaud and et al 2015). This is applied in order to have a higher representation of the features learned at a node level. It results in features for each subgraph (video input). This is an essential step, reducing the size of the overall graph and propagating the learned features for each subgraph representing a video input.

Figure 2 shows the main architecture for our SynchGraph, which stacks up multiple convolution layers. We have used a deeper architecture compared to the one introduced in (Kipf and Welling 2016) by adding a hidden layer. Having a deeper network helps in aggregating and translating the complex relationship between nodes to sub-graphs.

At each layer a GCN produces an output in the form of a feature matrix $Z_{N\times D}$, where D represents the dimension of output features for each graph and N is the number of nodes. Each layer can be represented by:

$$H^{(l+1)} = f(H^{(l)}, A), \tag{6}$$

$H^{(l)}$ represents the activation matrix at the lth layer and the activation matrix for the first layer is the feature matrix X. f is the propagation function that aggregates features at the lth layer with the adjacency matrix A, leading to features at the subsequent layer $l+1$.

Spectral graph convolution is applied to the graphs by applying Eigen-decomposition of the graph Laplacian. The

spectral convolutions are defined by the multiplication of graph signal $x \in \mathbb{R}^N$ (which is a scalar value for every node) with a filter $g_\theta = diag(\theta)$ where $\theta \in \mathbb{R}^N$ is in the Fourier domain (Kipf and Welling 2016). The spectral convolution can be translated by:

$$g_\theta * x = U_{g_\theta} U^T x \tag{7}$$

U represents the matrix of eigenvectors of the normalised graph Laplacian $L = I_N - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = U \Lambda U^T$, where Λ is the diagonal matrix of the eigenvalues. g_θ is a function of the eigenvalues of L. $U^T x$ is the graph Fourier transform of the graph signal x.

The input to the network consists of multiple sub-graphs each representing neural activities of a video input. The network consists of three layers followed by a pooling layer over graph (Duvenaud and et al 2015) in order to combine features from all sub-graphs and enable the classification of subgraph. The main learning model and propagation rule can be defined as follows:

$$Z = f(X, A) = \texttt{softmax}(\hat{A}\sigma(\hat{A}\sigma(\hat{A}XW^{(0))})W^{(1)})W^{(2)}), \tag{8}$$

where weights are defined by weights matrices with $W^{(0))}$ representing the input to hidden layer weight matrix, $W^{(1)}$ the weight matrix from hidden layer 1 to hidden layer 2 and $W^{(2)}$ is the hidden to output weight matrix. $\hat{A} = A + I_N$ is the adjacency matrix of the graph with added self connection and I_N is the identity matrix. The loss function is defined as the cross-entropy over labelled neurons:

$$\mathcal{L} = -\sum_{d \in y_D}^{.} \sum_{c=1}^{C} Y_{d,c} ln Z_{d,c} \tag{9}$$

y_D is a set of neurons that are labelled and C represents the dimension of the output classes; *i.e.*, six basic emotions. The networks weights $W^{(0)}$, $W^{(1)}$, and $W^{(2)}$ are trained with gradient descent, where the full training set is used in each iteration (Kipf and Welling 2016).

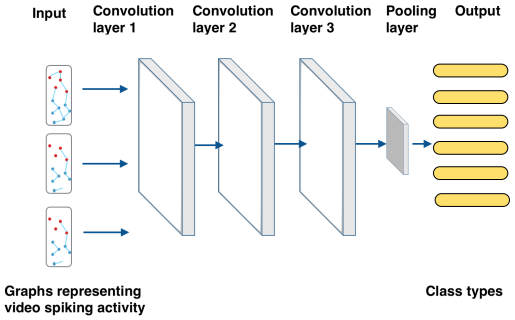

Figure 2: Architecture of Synch-Graph- – Graph Convolutional Network for Neural Synchrony

Experiments

In this section we evaluate the performance of Synch-Graph on two datasets for multisensory emotion recognition and compare with state-of-the-art techniques.

Datasets

We use two dataset to evaluate our model. The first dataset is the eNTERFACE'05 dataset (Pitas et al. 2006) with 42 participants composed of 81% male and 19% female participants. The audio is recorded at 48000HZ in 16 bit format. The second dataset is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018). The dataset consists of a balanced gender with 24 participants. The participants are actors reading a sentence in six different emotional states.

Implementation and Network Configuration

All experiments are implemented using Spiking neural simulator BRIAN (Goodman and Brette 2008). We use the same network architecture and parameters as (Diehl and Cook 2015), including input firing rates, membrane threshold and resting phase duration. Differently we add a convolutional layer in the excitatory layer for a better feature representation (Mansouri-Benssassi and Ye 2018), (Saunders and et al 2018). After feature extraction and transforming inputs into Poisson spike trains (Diehl and Cook 2015) for both audio and visual, we have set the number of neurons for each group to be proportionate to the dimension of inputs; that is, 40×388 for audio and 100×100 for visual neuron groups.

We have used a convolutional window for each modality with a convolutional window size of 40 and an initial number of features of 20 for each modality. Although setting features to a higher value and smaller convolutional window would increase the accuracy, we have chosen the above setting due to computational power limitations. The audio input is fed to the network after a 5ms delay. This is to model the natural temporal lag between visual and auditory sensory inputs in the brain. Recurrent connections between modalities are applied at the excitatory layer. This enables the cross-talk between audio and visual modalities and help simulate multisensory interaction where modalities influence each other during the learning process.

The constructed neural synchrony graph on RAVDESS dataset consists of 814 sub-graphs and 130008 nodes in total. On the eNTERFACE'05 dataset we have obtained 1260 sub-graphs and 201600 nodes in total. After obtaining the basic structure for each graph we prepare the input for the GCN. We have trained a three-layer GCN with a semi-supervised learning and have initialised the weights randomly (Kipf and Welling 2016). We use Adam optimisation and a learning rate of 0.0001. These hyper-parameters are chosen after experimenting with various learning rate starting from 0.01. We use hidden layers of 64 units in the second and third layer. We train the network for 500 epochs with a dropout rate of 0.5. We randomly shuffle the data and use a dataset split by 60% for training 20% for validation and 20% for testing.

Results and Discussion

Figure 4 shows the learning curve of Synch-Graph on training and validation data. The loss decreases after around 200 epochs and then stabilises. The validation loss becomes lower than the training at around 100 epochs. The figure also

Table 1: Comparison of accuracy for multisensory emotion recognition on RAVDESS

Technique	Feature Extraction	Fusion Method	Accuracy (%)
(Beard et al. 2018)	COVAREP,OpenFace	LSTM+GCA(Global Conceptualised Attention)	58.33
(Mansouri-Benssassi and Ye 2019)	LoG, MFCCs	Early cross-modal enhancement+SNN	83.3
(Alshamsi et al. 2018)	SVM +Early Fusion	Early Fusion with SVM	97.26
Synch-Graph	**LoG,MFCCs, SNN**	**Neural Synchrony with GCN**	98.3

Table 2: Comparison of accuracy in multisensory emotion recognition on eNTERFACE'05

Technique	Feature Extraction	Fusion Method	Accuracy (%)
(Beard et al. 2018)	Facial Landmarks	Early feature fusion	62.8
(Di Nardo, Pet- rosino, and Ullah 2018)	CNN	3D pyramidal neural network	71.47
(Fonnegra and Diaz 2018)	CNN and RNN	MLP	81.84
(Zhang et al. 2017)	CNN and 3DCNN	DBN	85.87
(Mansouri- Benssassi and Ye 2019)	LoG, MFCCs	Early cross-modal enhancement+SNN	86.3
Synch-GCN	**LoG,MFCCs, SNN**	**Neural Synchrony with GCNN**	96.82

shows that there is a small gap between the training and validation loss. Compared to the original architecture with two layers from (Kipf and Welling 2016) in Figure 3, we notice that the gap of the loss is bigger between validation and training. This shows that using only 2 layers needs more training epochs and training data. Having a third layer increases the learning capacity of the network.

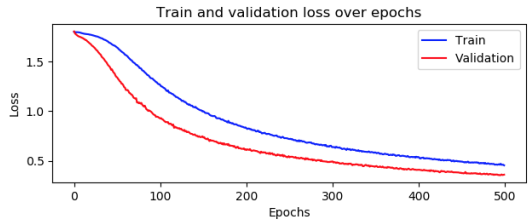

Figure 3: Loss on 2-layer GCN with RAVDESS dataset

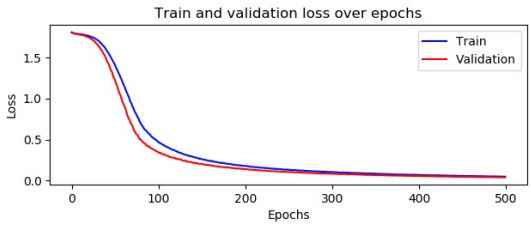

Figure 4: Loss on 3-layer GCN with RAVDESS dataset

Table 1 and 2 show accuracy comparison between our approach and the state-of-the-art techniques. Our approach has achieved an overall accuracy of 98.3% and 96.82% for RAVDESS and eNTERFACE'05 datasets respectively. This is significantly higher than state-of-the-art multisensory emotion recognition method tested on the same datasets.

The best performing state-of-the-art on the eNTERFACE'05 is the work presented by Zhang et al. (Zhang et al. 2017), which has achieved 85.97%. They have used DBN to learn feature representations on concatenated features from both audio and visual signals.

In comparison, our approach extracts information from both modalities and learning happens simultaneously between audio and visual inputs. Each modality influences the other during the learning process using connections between them. SNN permits to capture the multisensory learning through connections between audio and visual neuron groups. GCN helps to model and learn synchrony patterns of neuron groups and enable multisensory emotion recognition across them.

Because of this strength, our model outperforms the DBN approach by 10.85%. Figure 5 presents the comparison on individual classes. The DBN approach achieves a lower accuracy of 80% on three classes: sadness, fear, and surprise, while Synch-Graph has achieved a consistently high accuracy of $\geq$90% on all the classes. We also compare our method with a biologically inspired model with early cross-modal enhancement (Mansouri-Benssassi and Ye 2019). Our model outperforms it by 24% for RAVDESS dataset and 13.52% for eNTERFACE'05 dataset.

Figure 6 and 7 present the confusion matrices of both datasets. We can observe a balanced accuracy for all classes with a lowest accuracy of 92.9% for surprise class on RAVDESS dataset and 92% for sad class on eNTERFACE'05 dataset.

In addition we run ablation analysis on our proposed model by comparing it to unisensory and multisensory enhancement using SNN using the same datasets as summarised in Table 3. We have trained and run unisensory SNN for both RAVDESS and eNTERFACE'05 datasets with the

Table 3: Ablation analysis of Synch-Graph with unimodal and early cross-modal enhancement techniques

Modality	Feature extraction	Technique	eNTERFACE'05 (%)	RAVDESS (%)
Video	LoG	SNN	65.3	57.5
Audio	MFCCs	SNN	43.51	42.6
Video + Audio	LoG,MFCCs,SNN	Early cross-modal enhancement	86.3	83.3
Video + Audio	**LoG,MFCCs, SNN**	**Neural Synchrony with GCNN**	96.82	98.3

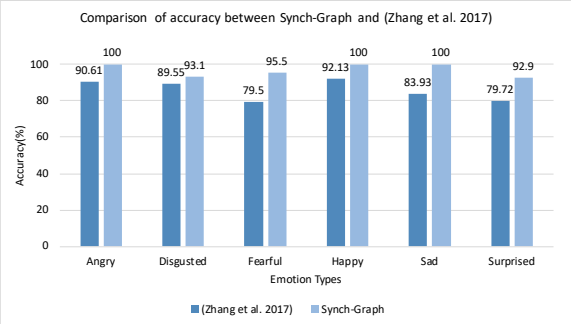

Figure 5: Comparison of accuracy by class type between state-of-the-art and Synch-Graph on eNTERFACE'05

same parameters and architecture as in (Mansouri-Benssassi and Ye 2018). We have run two separate SNNs for audio and visual data. The accuracy gain of Synch-Graph is over 50% compared to unisensory models and 10 $\sim$ 15% compared to early enhancement technique. This significant improvement in accuracy demonstrates the advantage of modelling and learning connections between neuron groups in multisensory emotion recognition.

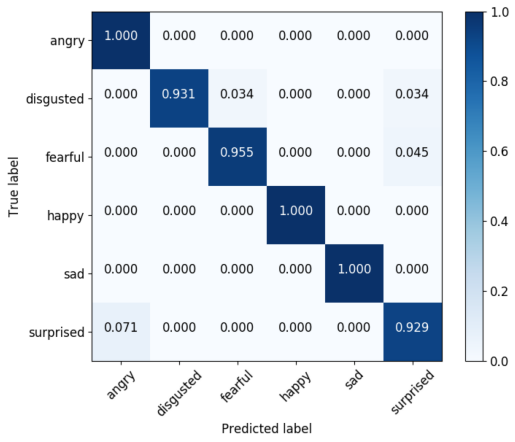

Figure 6: Confusion matrix for RAVDESS dataset

Conclusion and Future Work

In this paper, we present a bio-inspired approach for multisensory emotion recognition based on neural synchrony and

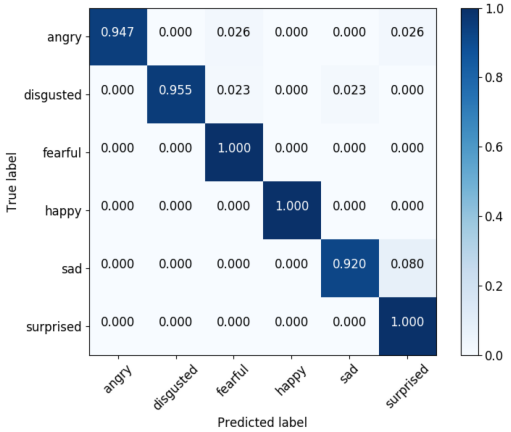

Figure 7: Confusion matrix for eNTERFACE'05 dataset

graph convolutional networks (GCN). Exploiting SNN with unsupervised STDP learning, temporal neural synchrony and the effectiveness of GCNs enables better feature representation and multisensory interactions modelling. More specifically,

- SNN with STDP learning enables features learning and cross-talk between both modalities.
- Computing with neural synchrony with spike timing and stimuli enables the integration of audio and visual data.
- GCN has demonstrated as a viable choice for modelling neuron activities and their interactions to facilitate learning complex patterns.

Our approach successfully translates the cross-modal talk and relation between audio and visual signals by using SNN representation of multisensory interaction. Using SNN to represent multisensory data can also alleviate the heterogeneity challenge of multisensory data. This is achieved by unifying all modalities features into a uniform input type – Poisson spike trains. In addition, representing data in graph addresses the fusion challenge by enabling data fusion while keeping the temporal and spatial relationship. Our approach will be particularly useful for robust in-the-wild emotion recognition, where there is uncertainty in either of the modalities. The neural synchrony patterns can help to enhance the recognition accuracy. Finally, this work paves a path to new opportunities in multisensory learning and integration field. Our future work will validate the robustness of Synch-Graph on in-the-wild datasets.

References

Balconi, M., and Carrera, A. 2011. Cross-modal integration of emotional face and voice in congruous and incongruous pairs: the p2 erp effect. *Journal of Cognitive Psychology* 23(1):132–139.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2):423–443.

Benssassi, E.; Gomez, J.-C.; Boyd, L.; Hayes, G.; and Ye, J. 2018. Wearable assistive technologies for autism: Opportunities and challenges. *IEEE Pervasive Computing*.

Brette, R. 2012. Computing with neural synchrony. *PLoS computational biology* 8(6):e1002561.

Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv:1312.6203*.

Chao, L.; Tao, J.; Yang, M.; Li, Y.; and Wen, Z. 2016. Long short term memory recurrent neural network based encoding method for emotion recognition in video. In *ICASSP '16*, 2752–2756. IEEE.

Diehl, P., and Cook, M. 2015. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience* 9:99.

Duvenaud, D. K., and et al. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, 2224–2232.

Felipe, C.; Luis J, M.; and Pedro, N. 2015. A novel multimodal emotion recognition approach for affective human robot interaction. In *ICIRS '15*.

Gao, H.; Wang, Z.; and Ji, S. 2018. Large-scale learnable graph convolutional networks. In *SIGKDD*, 1416–1424.

Garrido-Vásquez, P.; Pell, M. D.; Paulmann, S.; and Kotz, S. A. 2018. Dynamic facial expressions prime the processing of emotional prosody. *Frontiers in human neuroscience* 12:244.

Goodman, D., and Brette, R. 2008. Brian: a simulator for spiking neural networks in python. *Frontiers in Neuroinformatics* 2:5.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Representation learning on graphs: Methods and applications. *arXiv:1709.05584*.

Hazan, H.; Saunders, D.; Sanghavi, D. T.; Siegelmann, H.; and Kozma, R. 2018. Unsupervised learning with self-organizing spiking neural networks. In *IJCNN '18*, 1–6.

Henaff, M.; Bruna, J.; and LeCun, Y. 2015. Deep convolutional networks on graph-structured data. *arXiv:1506.05163*.

Jose, J. T.; Amudha, J.; and Sanjay, G. 2015. A survey on spiking neural networks in image processing. In *Advances in Intelligent Informatics*, 107–115.

Keil, J., and Senkowski, D. 2018. Neural oscillations orchestrate multisensory processing. *The Neuroscientist* 24(6):609–626.

Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*.

Livingstone, S. R., and Russo, F. A. 2018. The ryerson audio-visual database of emotional speech and song (ravdess). *PloS one* 13(5):e0196391.

Mansouri-Benssassi, E., and Ye, J. 2018. Bio-inspired spiking neural networks for facial expression recognition: Generalisation investigation. In *TPNC*, 426–437. Springer.

Mansouri-Benssassi, E., and Ye, J. 2019. Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks. In *IJCNN '19*.

Nian, F.; Chen, X.; Yang, S.; and Lv, G. 2019. Facial attribute recognition with feature decoupling and graph convolutional networks. *IEEE Access*.

Pitas, I.; Kotsia, I.; Martin, O.; and Macq, B. 2006. The enterface'05 audio-visual emotion database. In *ICDEW '06*.

Rathi, N., and Roy, K. 2018. Stdp-based unsupervised multimodal learning with cross-modal processing in spiking neural network. *IEEE Transactions on Emerging Topics in Computational Intelligence* 1–11.

Saunders, D. J., and et al. 2018. Stdp learning of image patches with convolutional spiking neural networks.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*.

Stein, B. E. 2012. *The new handbook of multisensory processing*. MIT Press.

Symons, A. E., and et al. 2016. The functional role of neural oscillations in non-verbal emotional communication. *Frontiers in Human Neuroscience* 10:239.

Wagner, J., and André, E. 2018. Real-time sensing of affect and social signals in a multimodal framework: a practical approach. *The Handbook of Multimodal-Multisensor Interfaces* 227–261.

Wu, Z., and et al. 2019. A comprehensive survey on graph neural networks. *arXiv:1901.00596*.

Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; and Tian, Q. 2017. Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 28(10):3030–3043.

Zhang, M.; Liang, Y.; and Ma, H. 2019. Context-aware affective graph reasoning for emotion recognition. In *ICME '19*, 151–156. IEEE.